Virology 466-467 (2014) 60–70

Contents lists available at ScienceDirect

Virology

journal homepage: www.elsevier.com/locate/yviro

Genome of brown tide virus (AaV), the little giant of the Megaviridae, elucidates NCLDV genome expansion and host–virus coevolution

Mohammad Moniruzzaman [a], Gary R. LeCleir [a], Christopher M. Brown [b], Christopher J. Gobler [c], Kay D. Bidle [b], William H. Wilson [d], Steven W. Wilhelm [a,*]

[a] Department of Microbiology, The University of Tennessee, TN 37996, United States
[b] Institute of Marine and Coastal Sciences, Rutgers, NJ 08901, United States
[c] School of Marine and Atmospheric Sciences, Stony Brook, NY 11794, United States
[d] Bigelow Lab for Ocean Sciences, Boothbay, ME 04544, United States

ARTICLE INFO

Available online 14 July 2014

Keywords:
NCLDVs
Megaviridae
Brown tides
Viral evolution
Giant virus
AaV
Algal virus

ABSTRACT

Aureococcus anophagefferens causes economically and ecologically destructive "brown tides" in the United States, China and South Africa. Here we report the 370,920 bp genomic sequence of AaV, a virus capable of infecting and lysing *A. anophagefferens*. AaV is a member of the nucleocytoplasmic large DNA virus (NCLDV) group, harboring 377 putative coding sequences and 8 tRNAs. Despite being an algal virus, AaV shows no phylogenetic affinity to the Phycodnaviridae family, to which most algae-infecting viruses belong. Core gene phylogenies, shared gene content and genome-wide similarities suggest AaV is the smallest member of the emerging clade "Megaviridae". The genomic architecture of AaV demonstrates that the ancestral virus had an even smaller genome, which expanded through gene duplication and assimilation of genes from diverse sources including the host itself – some of which probably modulate important host processes. AaV also harbors a number of genes exclusive to phycodnaviruses – reinforcing the hypothesis that Phycodna- and Mimiviridae share a common ancestor.

Introduction

The presence of viruses with large genomes in marine systems has stimulated questions concerning the origin and phylogenetic histories of these unique particles. Such viruses have been found to infect diverse hosts, including eukaryotic phytoplankton (Fitzgerald et al., 2007; Moreau et al., 2010; Santini et al., 2013; Wilson et al., 2005), non-photosynthetic protists like *Acanthamoeba* (includes Mimivirus (Raoult et al., 2004), Marseillevirus (Boyer et al., 2009) and most recently Pithovirus (Legendre et al., 2014)) and the zooplankton *Cafeteria roenbergensis* (Fischer et al., 2010). Genomic characterizations of these viruses have dramatically altered our perceptions of the breadth of functional potential for virus particles. Each of these viruses is characterized by a large genome size, spanning several hundred to thousands of kilobases, and an exceptionally diverse gene content that is atypical of most other viruses. Analyses of the genome architecture of these viral "leviathans" has revealed patterns of massive gene duplication (Suhre, 2005) and gene acquisition from diverse sources, including their putative hosts (Filee et al., 2008). With the discovery of large-genome viruses, established phylogenetic classifications of NCLDVs infecting unicellular eukaryotes NCLDVs have been challenged, with proposals to reclassify NCLDVs having the largest genomes in the 'Megaviridae' clade irrespective of their host range (Santini et al., 2013). Emerging information on the genomic sequence and architecture of these viruses has the potential to redefine our understanding of virus function and evolution, including kindling debate that these viruses may represent a new, fourth domain of life (Boyer et al., 2010; Williams et al., 2011).

Aureococcus anophagefferens, a unicellular microalga, is a pelagophyte which causes recurrent brown tide blooms in the coastal and estuarine waters of the eastern United States, South Africa (Gobler and Sunda, 2012) and China (Zhang et al., 2012). The economic and ecological effects of brown tides are significant; blooms cause severe light attenuation in the coastal waters, resulting in destruction of sea grass beds (Gobler and Sunda, 2012), an important nursery for marine life. Brown tides are also toxic to bivalves and have contributed towards the collapse of multiple shellfisheries (Gobler and Sunda, 2012). Early transmission electron micrographs of *A. anophagefferens* alluded to the importance of viruses in the ecology of this organism, as they

* Corresponding author.
E-mail address: wilhelm@utk.edu (S.W. Wilhelm).

http://dx.doi.org/10.1016/j.virol.2014.06.031
0042-6822/

revealed virus-like particles in natural populations (Sieburth et al., 1988). Subsequent studies demonstrated that viruses likely played an important role in the modulation of bloom events: Gastrich et al. (2004) found that up to 37.5% of the population of *A. anophagefferens* may be visibly infected during bloom peak (Gastrich et al., 2004), suggesting this virus may be present at total abundances as high as $\sim 10^{20}$ particles in Great South Bay (NY).

Initial attempts at isolation of *A. anophagefferens*-specific viruses reported that phage-like tailed particles were present as lytic agents of cells in cultures (Garry et al., 1998; Milligan and Cosper, 1994). However, subsequent research identified and isolated a large, icosahedral virus with a diameter of ~140 nm that was morphologically consistent with the earlier observations from blooms of *A. anophagefferens* (Gastrich et al., 2002; Rowe et al., 2008). Several aspects of viral infection dynamics of *A. anophagefferens* have already been investigated (Gastrich et al., 2004; Gobler et al., 2007), but a crucial step in understanding the molecular mechanisms of host–virus interactions and indeed the ecology of giant viruses is to decipher genomic information. The 56 Mbp genome of host, *A. anophagefferens*, has recently been described (Gobler et al., 2011). We now report the complete genome sequence of the *A. anophagefferens* virus (AaV) which we assembled from combined *Illumina*™ and *454*™ pyrosequencing data. We provide a comprehensive analysis of the gene content, genome architecture and phylogenetic position of AaV.

Results and discussion

General genome features

A. anophagefferens Virus (AaV) has a linear double-stranded DNA genome with a size of 370,920 bp (Fig. S1). Applying a conservative annotation process (SI Appendix A), we identified 377 putative coding sequences in the genome of AaV, with a coding density of 88.3%. Such a high coding density is typical of large dsDNA viruses. The genome is A+T rich with a G+C content of 28.7%, in stark contrast to the host, which has a very high G+C content (69.5%) (Gobler et al., 2011). AT richness of the genome is also reflected in the codon usage of AaV; ~25% of the codons contain only A or T, whereas ~41% of the codons contain at least two A or T (excluding the stop codons). Putative coding sequence (CDS) length ranges from 52 to 2076 (with an average of ~290) amino acids.

Among the putative coding regions, 53% (200 of the 377) have significant (*E*-value < 1e-05) sequence similarity to proteins in the NCBI *nr* database, with 67 best matches to nucleocytoplasmic large DNA viruses (NCLDVs), 56 with eukaryotes, 72 with bacteria/bacteriophage and 5 best matches to archaea (Fig. 1a). Twenty eight of the CDSs having best match to NCLDVs have no homologs in the three domains of life (NCLDV specific hypothetical CDSs). Six AaV CDSs were most similar to sequences from the host (in terms of best BLASTp hit) *A. anophagefferens,* indicating the possibility of horizontal gene transfer (HGT) between the host and the virus (Table S2). Among the 177 CDSs having no matches in the NCBI nr database, 34 had significant hits to NCBI environmental database (env_nr). Eighty seven (23%) of the AaV CDSs could be assigned within the Cluster of Orthologous Group categories (COG) (Supplemental Figs. S1 and S2), providing insight into their potential function.

A gene complement typical of NCLDVs

AaV harbors 9 of the 10 genes that are present in most members of all the NCLDV lineages defined by Yutin et al. (2009) (Supplemental Table S1), suggesting that AaV belongs to this group. Most NCLDVs characterized to date harbor genes involved in the basic biochemistry of "living" (i.e., cellular) organisms: replication, transcription and translation. At least 18 genes within AaV can be categorized in the NCVOG (Nucleocytoplasmic Virus Orthologous Group) category 'Transcription and RNA processing' (Supplemental Figs. S1 and S2). AaV has two copies of RNA polymerase II large subunit (Rpb1) (AaV_242, AaV_320), possibly as a result of paralogous expansion (Supplemental Table S3), and two non-paralogous copies of RNA polymerase II small subunits (Rpb2) (AaV_222, AaV_370). Such a phenomenon has also been reported in *Phaeocystis globosa* virus 16T (Santini et al., 2013), which shows gene duplication of Rpb2 subunits. AaV also harbors six other eukaryotic RNA polymerase subunits. Proteins involved in transcription initiation and elongation, namely Transcription initiation factor TFIIB (AaV_203), TATA-box binding protein (AaV_117) and a transcription elongation factor TFIIS (AaV_381), are present in the AaV genome. The acetylation state of histones in the chromatin is an important regulator of transcription in eukaryotic organisms (Marmorstein and Roth, 2001). ELP3-histone acetyl transferase genes are present in the *C. roenbergensis* virus (Fischer et al., 2010) and the *P. globosa* virus genome. AaV also harbors this gene (AaV_368), and phylogenetic analysis suggests a eukaryotic origin for this gene (Supplemental Fig. S3a). The maintenance of this gene in viruses infecting hosts with diverse lifestyles indicate that it was possibly present in the common ancestor of these viruses and may have an important function in modulating the transcriptional state of the host or virus maturation.

Although members of the NCLDV family are mostly independent of the host for replication and transcription, they typically depend on host protein synthesis machinery (Koonin and Yutin, 2010). However, genes coding for tRNAs and proteins involved in translation have been found in large viruses (Raoult et al., 2004; Wilson et al., 2005). Consistent with this, we identified eight tRNA genes in AaV (Supplemental Table S4, Fig. S1). Among these, five are present as a cluster starting at position 322,252 bp. Elongation factor 5A (AaV_110) and eIF 1α (AaV_118), genes involved in translation elongation and initiation, respectively, also occur in AaV. Translation elongation factor 5A is unique to AaV among large viruses (Supplemental Fig. S3b).

While elevated photosynthetically active radiation has been shown to accelerate the virus-mediated lysis of the host (Gobler et al., 2007), higher intensity UV radiation can introduce DNA damage, including pyrimidine dimers (Weinbauer et al., 1997). Photolyase genes have been found in a number of large viruses (e.g., Fischer et al., 2010) and their role in repairing pyrimidine dimers in a NCLDV has been demonstrated (Srinivasan et al., 2001). A class II photolyase (AaV_082) in AaV likely plays a similar role. A lambda-type exonuclease (AaV_159), a Holliday junction resolvase (AaV_201) and a dUTPase gene (AaV_318) are all present in AaV and relevant given the high A+T content in its genome. A MutS7 gene, putatively involved in mismatch repair, has been found in all members of the *Mimiviridae* family (Ogata et al., 2011), and is also present in AaV. The MutT gene (nudix hydrolase) is involved in preventing the mis-incorporation of dGMP and thus transversion mutations (Akiyama et al., 1989) and there are two copies in the AaV genome (AaV_234, AaV_173).

Enzymes involved in ubiquitination are found in all NCLDV lineages. AaV encodes five E3 ubiquitin ligases, one E2 ubiquitin ligase, one POZ domain protein (part of the SCF–E3 complex) and also a Ulp1 family thiol protease, a deubiquitination protein. This arsenal of proteins likely contributes toward its ability to overcome the host's defense against viral infection by interfering with Ub signaling (Iyer et al., 2006).

Unique CDSs derived from the host and other sources

It has been established that host-derived genes in cyanophage and large DNA viruses play key roles in resource acquisition by

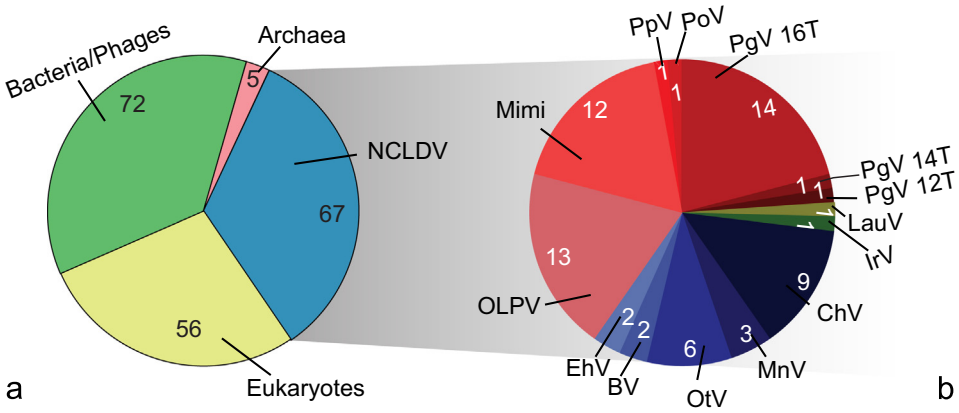

Fig. 1. Best BLASTp hits of AaV proteome against (a) NCLDVs and three domains of life. (b) Different NCLDVs. IrV—Iridoviridae, PpV—*Phaeocystis pouchetii* virus, PoV—*Pyramimonas orientalis* virus, PgV16T—*Phaeocystis globosa virus* 16T, PgV 14T—*P. globosa virus* 14T, PgV 12T—*P. globosa virus* 12T, ChV—*Chlorella* viruses, MnV—*Micromonas* viruses, OtV—*Ostreococcus tauri* viruses, BV—*Bathycoccus* sp. virus, EhV—*Emiliania huxleyi* viruses, OLPV—Organic Lake Phycodnaviruses, Mimi—Mimiviridae, LauV—Lausannevirus. Hits to the Megaviridae clade are presented as shades of red whereas phycodnavirus hits are in shades of blue. Irido- and Lausannevirus hits are in shades of green.

infected hosts and facilitate viral synthesis and host lysis (Hill, 2006; Monier et al., 2009; Wilson et al., 2005). For a number of AaV CDSs, no homologs could be identified in other NCLDVs (Supplemental Table S5). While BLASTp identified only six genes with best matches to host proteins, our phylogenetic reconstructions imply that at least 13 genes were possibly acquired from the host. (Supplemental Table S5, Fig. S3). Among these, glucuronyl hydrolase (AaV_078) (Supplemental Fig. S3c) and pectate lyases (AaV_003, 375, 038) (Supplemental Fig. S3d) are both present in the host alga (Gobler et al., 2011). Brown tides are associated with severe light attenuation, which minimizes light available for photosynthesis (Gobler et al., 2011; Gobler and Sunda, 2012). As such, glucuronyl hydrolase and pectate lyase genes in AaV may permit *A. anophagefferens* to derive energy from abundant sources of organic carbon during blooms, enhancing the ability of the infected cell to generate energy when light levels are reduced and as the chloroplast is degraded in the late stages of infection (Gastrich et al., 1998). Another putative host-derived gene is an intramembrane rhomboid family serine protease (AaV_077) (Supplemental Fig. S3e). Given that *A. anophagefferens* can use complex sources of organic nitrogen, this protease may enhance the ability of infected *A. anophagefferens* to access nitrogen (Gobler et al., 2011; Gobler and Sunda, 2012). Collectively, the presence of genes associated with the degradation of organic compounds by *A. anophagefferens* in AaV suggests they were obtained from the host and retained due to enhanced fitness they provide.

AaV harbors a calpain family thiol protease (AaV_045) not reported in any other NCLDVs (Supplemental Fig. S3f). Calpains have diverse functions, including cell cycle regulation and caspase-independent apoptosis (Nemova et al., 2010). Modulation of host cell apoptosis by NCLDVs, especially among the *Phycodnaviridae* (reviewed in Bidle and Vardi, 2011), is an emerging feature of host–virus interactions and appears to be a shared trait amongst diverse algal viruses (Bidle and Vardi, 2011). Studies have demonstrated a pivotal role for viral induction of the host programmed cell death machinery in the propagation of *Emiliania huxleyi* viruses (EhV) (Bidle et al., 2007; Vardi et al., 2009), with this trait being conserved among a mixed population of both host and EhV genotypes (Vardi et al., 2012). In other systems (e.g., Hepatitis C) activation of calpains have been shown to inhibit the host's extrinsic apoptotic signaling pathway, which is necessary for successful infection. (Simonin et al., 2009). Consequently, the possession of such a protease in the AaV genome might allow for avoidance of the host's virus exclusion strategy.

Among NCLDVs, ion channel proteins have only been reported in the *Phycodnaviridae*. Potassium channel proteins in *Chlorella* viruses (Thiel et al., 2011) are well-studied and critical in the infection process (Greiner et al., 2009; Neupartl et al., 2008). We located a putative potassium ion channel protein (AaV_153) with a length of 157 aa as well as a putative small conductance mechanosensitive ion channel protein (AaV_043). Small conductance mechanosensitive channels are implicated in counteracting the osmotic pressure inside the cells (Wilson et al., 2013). The presence of two ion channel proteins, which operate under two distinct stimuli (ionic and mechanical) in AaV raise the possibility that they might have important (and different) roles during different stages of infection.

AaV has two prenyl transferases, commonly involved in lipid metabolism. One of these, undecaprenyl pyrophosphate synthase, was putatively derived from the host (AaV_255) (Supplemental Fig. S3g). The second protein, UbiA prenyltransferase (AaV_373) (Supplemental Fig. S3h), is a key enzyme in the biosynthesis of Ubiquinone (Szkopinska, 2000), a critical molecule in the respiratory electron transport chain. The presence of UbiA along with an AIM24 domain protein (AaV_144) (Supplemental Fig. S3i), that has a role in mitochondrial biogenesis (Hess et al., 2009), imparts the potential for AaV to further modulate the host's energy generating processes.

Carbohydrates can mediate interactions in diverse virus-host systems. The genome of AaV harbors a carbohydrate sulfotransferase (AaV_102) (Supplemental Fig. S3j), a gene involved in producing sulfated carbohydrates. Heparan sulfate, a sulfated carbohydrate, is known to be a surface receptor for a number of viruses including Vaccinia and Herpes Simplex (Zhu et al., 2011). In the case of *E. huxleyi*, EhV-86 encodes C-type lectin-containing protein that associates with purified lipid rafts from 2 h post-infected host cells, arguing that EhV infection occurs at the interface between virus proteins and host lipid-raft sugar–lipid moieties (Rose et al., 2014). No studies have been conducted on the molecular mechanisms of AaV–host interactions yet, so whether sulfated carbohydrates have any role in such interaction remains an open question. The *A. anophagefferens* genome is highly enriched in sulfatase genes that encode proteins that degrade sulfonated polysaccharides (Gobler et al. 2011) which may assist in discouraging the attachment of AaV to its cell surface.

Finally, AaV encodes phaeophorbide *a* oxygenase (PaoA; AaV_372) (Supplemental Fig. S3k), which is a key enzyme in chlorophyll catabolism (Pruzinska et al., 2003). PaoA was also present in Organic

Lake Phycodnavirus 2 (Yau et al., 2011), a virus assembled using metagenomic data from a hypersaline lake in Antarctica. Occurrence of this gene in two viruses from distinct geographic locations suggests that it was probably present in the common ancestor and might play role in modulating the host cellular processes. However, the possibility of independent acquisition of the gene through HGT cannot be discounted, either.

Putative role(s) for repetitive DNA elements in AaV

Repetitive DNA elements occur frequently in large DNA virus genomes. For example, *C. roenbergensis* virus (CroV) and Mimivirus have FNIP repeats (Pfam: PF05725) (Fischer et al., 2010; Raoult et al., 2004) while three distinct families of repeats with no homology within available databases were reported in EhV-86 (Allen et al., 2006). Approximately 11.3% of the genome of AaV is comprised of a lysine-enriched domain of unknown function (DUF285, Pfam: 03382), which is distantly related to leucine rich repeats. DUF285 domain regions are sequestered in putative coding sequences, resulting in a large paralogous family of 50 ORFs (Supplemental Table S3, Fig. S1), which is 13.25% of the total gene content. These CDSs range from 101 amino acids to 708 amino acids in length and all of them contain ≥ 1 copy of either partial or complete DUF285 domain defined in the Pfam database. Phylogenetic analysis reveals that this sequence probably originated in a bacterium (Fig. S3l). Repeats characterized by DUF285 domains occur sporadically in unicellular microbes, especially in the obligate endosymbiotic class Mollicutes, and also in unicellular photosynthetic eukaryotes (Roske et al., 2010). ORFs containing these domains were termed Palindromic Amphipathic Repeat Coding Elements (PARCELs), characterized by repeating elements displaying dyad symmetry and variable hydrophilic and conserved hydrophobic regions (Roske et al., 2010). These ORFs are also found as part of some bacterial mobile elements and plasmids. It has been suggested that PARCELs have spread in diverse bacterial and eukaryotic lineages through HGT and intra-genomic shuffling (Roske et al., 2010). The sequence characteristics of PARCELs endow them with potential roles in gene expansion and recombination (Roske et al., 2010).

Twenty of these PARCELs are present as tandem repeats at the 5′ extremity of AaV genome (genome location **A**: 1908–22,906 bp, interrupted by three other CDSs), while another cluster of 9 genes are present near the other end of the genome (genome location **B**: 330,172–339,020 bp) (Fig. 2). Both clusters are on the positive strand while the rest of the PARCELs are evenly distributed on the negative strand (Fig. S4). Additionally, two distinct conserved domains (which we denote as Motif_A and Motif_B) have been found to be consistently present at the upstream regions of the positive and the negative strand PARCELs, respectively (Supplemental Fig. S5). In the genome of host *A. anophagefferens* we have identified occurrences of the DUF285 motif at 90 distinct loci distributed across 32 scaffolds. The presence of PARCELs in both the host and virus genomes is intriguing. Since phylogenetic analysis suggests the possible origin of host PARCELs in bacteria (Fig. S3l), the most parsimonious scenario is the mobilization of this sequence to AaV from the host (upon uptake from bacteria) and its subsequent intra-genomic duplication. Alternatively, it is possible that this sequence is in flux between the host and the virus, playing a role in host–virus coevolution. Lineage-specific gene expansion contributed to the genome growth of other NCLDVs substantially (Iyer et al., 2006), and a similar mechanism is probably in effect in the genome of AaV. Duplicated genes can also go through neo-functionalization, a process where the daughter copy assumes a new function distinct from the mother gene (Liu et al., 2011). In accordance with this mechanism, we have found one of the PARCELs (AaV_220) containing a U-box domain fused to it, which probably functions as an E3 ubiquitin ligase (Ohi et al., 2003). The presence of conserved motifs at the upstream of the PARCELs and their orientation in antiparallel directions (Supplemental Fig. S5) inside the genome suggest that they are subject to intra-genomic mobilization. PARCELs at the extremities of the genome (AaV_001 & AaV_382) are present as inverted repeats (Supplemental Fig. S4), which potentially mediate circularization of AaV genome as has been found in Mimivirus and some other NCLDVs (Raoult et al., 2004).

The phylogenetic position and evolutionary history of AaV

DNA polymerase gene-based phylogeny clustered AaV in the *Mimiviridae* family (Fig. 3). The only other algal virus that clusters in *Mimiviridae* family is *P. globosa* 16 T. Two viruses from metagenomes generated in a hypersaline Antarctic lake (Organic Lake Phycodnavirus 1 and 2) (Yau et al., 2011), having no known hosts, also cluster in the *Mimiviridae*. Based on common marker genes and large genome sizes, it has been hypothesized that these viruses have a common ancestor, and a new group called 'Megaviridae' has been proposed that further extends the *Mimiviridae* family and is independent of hosts (Santini et al., 2013). Sixty-seven AaV proteins have their highest sequence similarity to large DNA viruses, among which, 65% are to Megaviridae family (Fig. 1b). Recruitment of whole genomes of NCLDVs to AaV demonstrate more coverage from the Megaviridae members relative to other NCLDVs (Fig. 4), supporting our suggestion that AaV is more similar to Megaviridae than to phycodnaviruses. Yutin et al. (2013) grouped the proteins of all the Megaviridae members, generating the 'Mimivirus cluster of orthologous groups (MimiCOGs)' (Yutin et al., 2013). Fifty two MimiCOG family genes are commonly shared among this group and, despite having the smallest genome, AaV possesses 46 of these 'core' genes (Supplemental Table S1).

Apart from their large genome size, a key feature of the members of the Megaviridae group is the presence of both asparagine synthetase and MutS7 genes. Based on our analyses, it is evident that AaV belongs to this proposed clade; however, it lacks the asparagine synthetase gene, raising questions concerning the universality of this gene within the Megaviridae. Other traits consistent with other Megaviridae to date (the presence of a "virus factory" within infected cells as well as the presence of virophage) are also not evident in our observations to date: while these latter traits require deeper investigation (e.g., TEM observations of infected cells in an effort to define the presence of virus factories), the data we have gathered to date highlights the difficulty of defining 'core' genes or traits in any system (Kislyuk et al., 2011).

Based on synapomorphies, it has been suggested that Phycodna- and Mimiviruses originated from a common ancestor (Iyer et al., 2006). Indeed, in contrast to our conclusion that AaV belongs within the Megaviridae, 22 of the AaV genes had highest sequence similarity with phycodnaviruses, and in a number of cases to genes exclusively present in phycodnaviruses. Among these are a phosphate starvation-induced protein (AaV_210) (Supplemental Fig. S3m), a RNA polymerase sigma factor 70 (AaV_076, putatively host derived), two copies SCF ubiquitin ligase (AaV_357, AaV_123), a zinc finger domain protein (AaV_380) and two paralogs of laminin G domain-containing protein (AaV_024, AaV_386). A number of hypothetical genes exclusive to phycodnaviruses were also found in the AaV genome (Table S6). An intriguing finding was the presence of 4 paralogous copies of a phycodnavirus-specific hypothetical gene (Paralog group 17; Supplemental Table S3) in the NCVOG cluster 1343, hereinafter denoted as 'AaV and phycodnavirus-specific highly similar genetic element' (AP_HGE). This gene is also found in several strains of *Paramecium bursaria* – *Chlorella* virus in multiple copies (Yutin et al., 2013). In AaV, these elements share very high sequence homology, with average

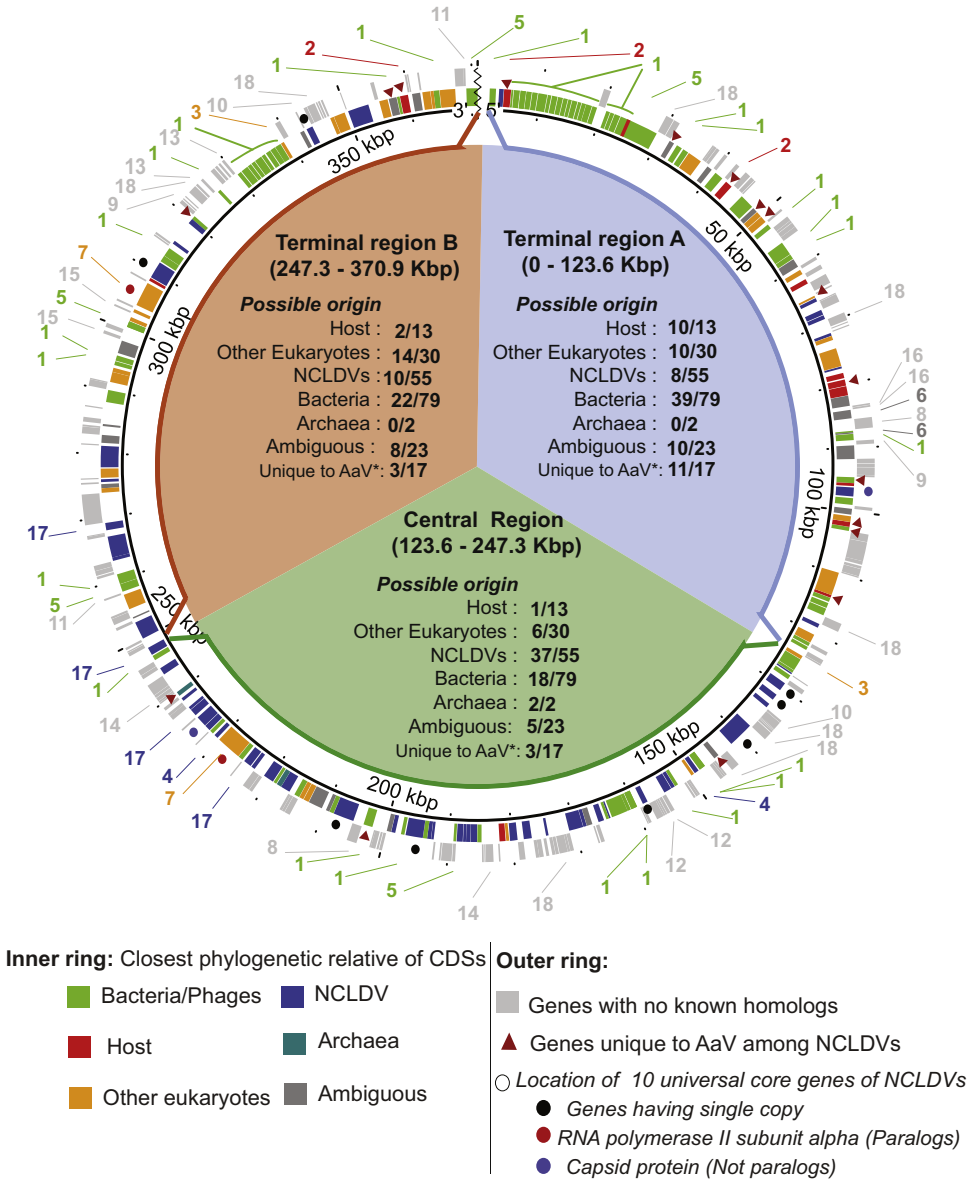

Fig. 2. Locations of the genes based on possible phylogenetic origins. To demonstrate the biased distribution of genes derived from different sources, the genome was divided into three equal sections to represent the central core and two terminal regions. Each of the sections are 123.6 kbp in length. The location of the paralogous genes are shown using the numbers that represent each of the paralog groups (SI Table 4). Genes with putative origin in host (red), other eukaryotes (orange), bacteria (green), archaea (teal) are shown. Genes for which the origin could not be inferred are depicted in gray (ambiguous origin). Locations of genes unique to AaV are marked with dark red triangles whereas universal NCLDV core genes are denoted by circles. The pie chart inside the genome map represents the three equal sections of the genome. Number of genes of different origins (Host, Other Eukaryotes, NCLDVs, Bacteria, Archaea and Ambiguous origin) located in each of the regions are presented inside the respective sections of the pie chart. For example, the central region harbors 32 of the 49 genes putatively derived from an NCLDV ancestor. ** Genes 'unique to AaV' refer to the genes that are only found in AaV, to date, among the NCLDVs.*

pairwise similarity of 90.5% and 80% at nucleotide and amino acid level, respectively. Furthermore, 90% of the 143 nucleotides immediately upstream and 82% of the 50 nucleotide positions immediately downstream of these ORFs are also fully conserved (Supplemental Fig. S6). Although AP_HGEs do not match known mobile genetic elements, the possibility that the conserved sequence signatures may contribute to the mobility of these ORFs cannot be ruled out.

Based on the similarity between AaV CDSs and proteins from other sources (Fig. 1a), we hypothesized that AaV has acquired genes from diverse sources. To reconstruct the evolutionary history of the AaV genes, we carried out a comprehensive maximum likelihood phylogenetic analyses for genes having homologs in diverse domains of life. According to the phylogenetic analyses, 78 genes possibly originated from bacteria, including the 50 genes in paralog group 1 (DUF285 domain containing proteins) (Supplemental Dataset 1, Table S3). Fifty-five genes showed highest phylogenetic affinity to NCLDVs. Eight genes were found to have closest similarity to the corresponding host protein (Fig. S3b, c, e and g, Table S2, Supplemental Dataset 1), indicating a relatively recent horizontal gene transfer. Thirty one genes appear to be acquired from eukaryotes other than the host. Another five proteins clustered with bacteria and the host (as the only eukaryote), indicating a history of gene transfer among bacteria, *Aureococcus* and AaV. Finally two genes possibly originated from archaea as suggested by our phylogenetic analysis. Phylogenetic reconstructions of these CDSs are available in the Supplemental dataset 1 and Fig. S3a–m.

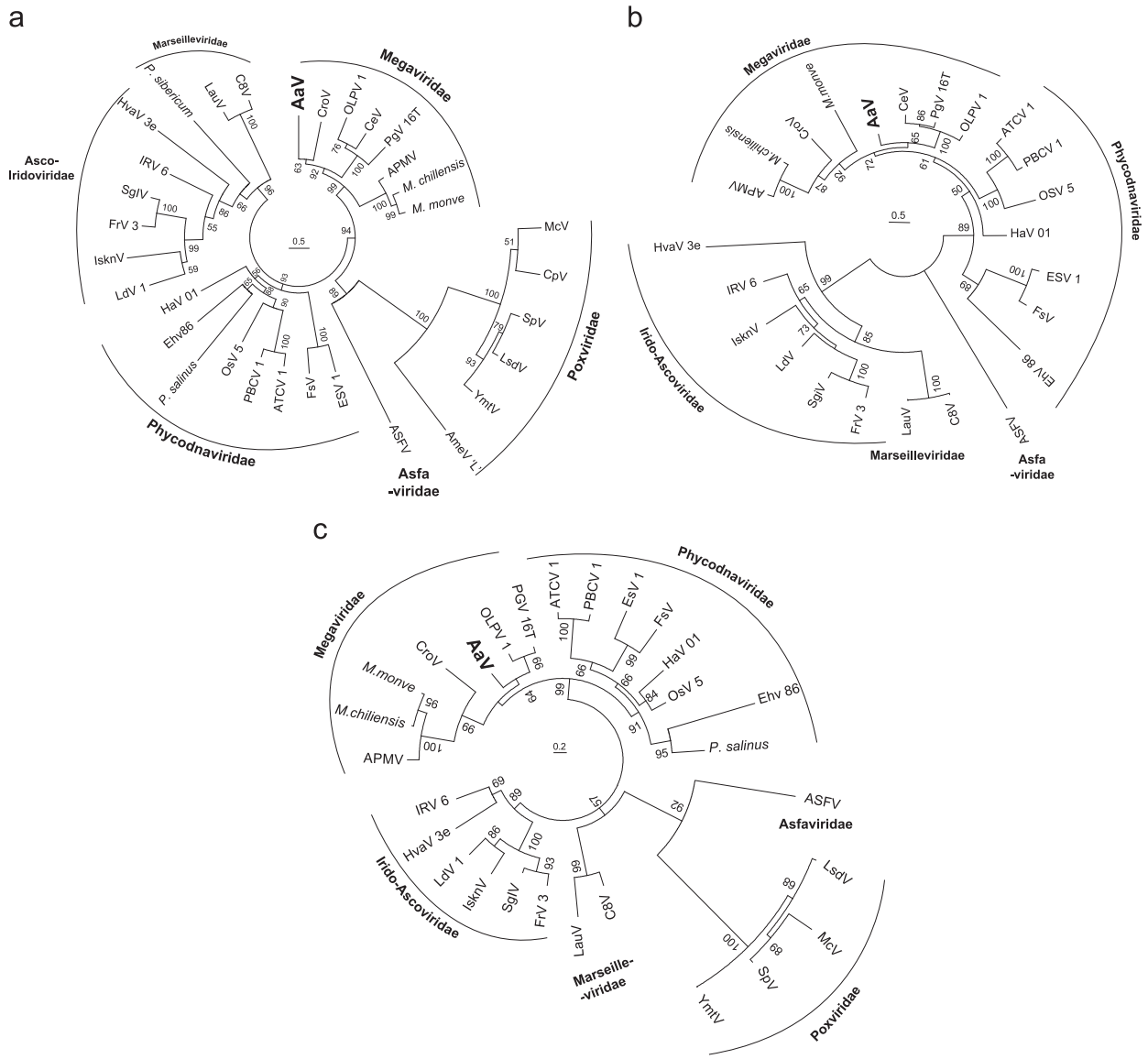

Fig. 3. Maximum likelihood phylogenetic trees of (a) B family DNA polymerase (AaV_141), (b) Major capsid protein (AaV_096) and (c) A32-like virion packaging ATPase (AaV_165) with other NCLDV members. The capsid homologs from the Poxviridae family are highly divergent and were not included in the major capsid protein phylogenetic analysis. The trees were constructed from amino acid alignments of the respective proteins. The Expected-Likelihood Weights (ELW) of 1,000 local rearrangements were used as confidence values for the nodes. The abbreviations are as follows: AaV, *Aureococcus anophagefferens* virus; CroV, *C. roenbergensis* virus; OLPV 1, Organic Lake Phycodnavirus 1; CeV, *Chrysochromulina ericina* virus; PGV 16T, *Phaeocystis globosa* virus 16T; APMV, *Acanthamoeba polyphaga* mimivirus; *M. chilensis*, *Megavirus chilensis*; *M. monve*, *Moumouvirus monve*; McV, *Mollascum contagiosum* virus; CpV, Canarypox virus; SpV, Swinepox virus; LsdV, Lumpy skin disease virus; YmtV, Yaba monkey tumor virus); AmeV 'L', *Amsacta moorei* entomopoxvirus L; ASFV, African swine fever virus; ESV 1, *Ectocarpus siliculosus* virus 1; FsV, *Feldmannia* species virus; ATCV 1, *Acanthocystis turfacea* Chlorella virus 1; OsV5, *Ostreococcus* virus OsV5; *P. salinus*, *Pandoravirus salinus*; EhV-86, *Emiliania huxleyi* virus 86; Hav 01, *Heterosigma akashiwo* virus 01; LdV 1, Lymphocystis disease virus 1; IsknV, Infectious spleen and kidney necrosis virus; Frv 3, Frog virus 3; SgIV, Singapore grouper iridovirus; IRV 6, invertebrate iridescent virus 6; HvaV, *Heliothis virescens* ascovirus 3e; *P. sibericum*, *Pithovirus sibericum*; LauV, Lausannevirus; C8V, Cannes 8 virus.

The genomic architecture of AaV

An initial observation, that most genes having phylogenetic affinity to the host proteins are located at the terminal regions of the genome, prompted us to further investigate the spatial distribution of the CDSs in the AaV genome. We partitioned the genome map into three sections of near-equal length (123.6 kbp starting from 5′ end (Terminal region A), 123.6–247.3 kbp (Central region) and 247.3–370 kbp (Terminal region B); Fig. 2). Terminal region A contains 10 of the 13 genes having phylogenetic affinity to the host, with three in tandem (AaV_076, 077 and 078), suggesting the concurrent acquisition of these genes from the host. Twenty-five of the 31 genes having origin in other eukaryotes are also situated within the two terminal regions. Interestingly, of the 17 genes that are unique to AaV among the NCLDVs (Supplemental Table S5), 14 are found in these two terminal sections. Genes unique to AaV are unlikely to have been vertically inherited from the ancestral virus and were possibly acquired through HGT. Sixty of the 78 genes putatively derived from bacteria are distributed in the terminal regions, most of which are paralogous copies of DUF285 domain-containing ORFs. One of the most interesting observations was the presence of seven universal NCLDV specific 'core' genes (Yutin et al., 2009) in the central region along with a capsid protein (AaV_247). This region also harbors 20 of the 28 NCLDV specific hypothetical CDSs that are found in AaV (Supplemental Table S6). In total, the central region accommodates 37 of the 55 AaV genes putatively having origin in NCLDVs. Taken together, these observations suggest that the

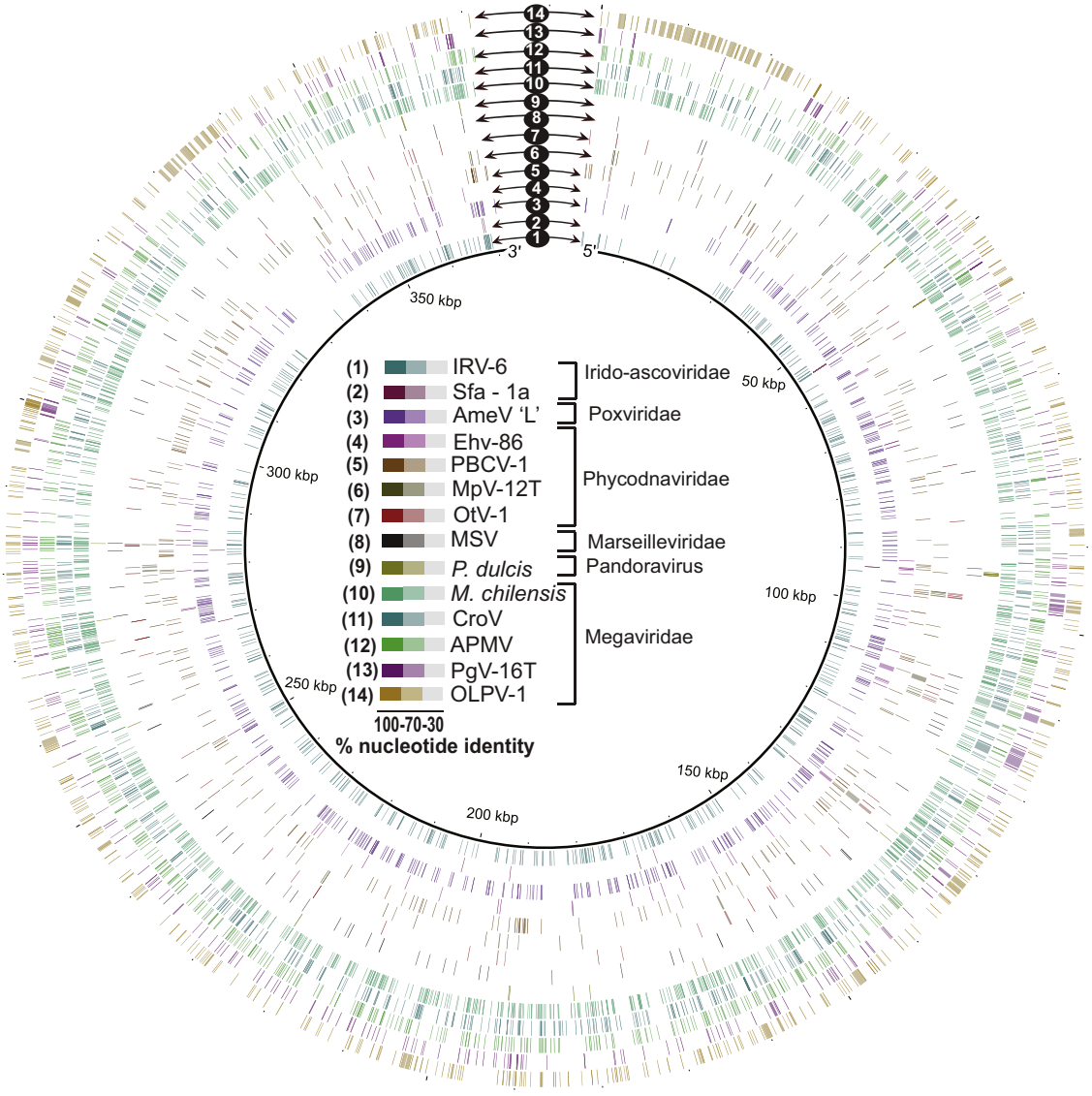

Fig. 4. BLASTn hits of whole genomes of NCLDV members recruited against the genome of AaV. The pairwise alignments (*E*-value < 1-e05) that each of the compared genomes produced with that of AaV were mapped on the AaV genome. The density of colors on each ring represents the identity percentage of nucleotides shared by a particular virus with AaV. This figure illustrates the fact that AaV genome shares more regions of sequence similarity to the Megaviridae family members than to members of other NCLDV families.

ancestral version of AaV was probably a much smaller NCLDV, and that this virus has expanded its genome by accruing genes at the terminal regions from diverse sources including the host itself.

It has been proposed that the Mimivirus genome evolved to its "giant" size by accumulating genes from host and other eukaryotic organisms at the terminal regions of the genome (Filee et al., 2008) and acquisition of bacterial genes by lineages of NCLDVs has also been demonstrated (Filee et al., 2007). Being a member of the Megaviridae clade, AaV shows a similar pattern of gene acquisition. In contrast to Mimivirus, however, AaV is unlikely to have intimate contact with bacteria inside the host, since *A. anophagefferens* is not known to be phagotrophic (Gobler and Sunda, 2012). *A. anophagefferens* can degrade foreign organic matter to derive energy and thus likely comes into frequent contact with foreign nucleic acids (Gobler et al., 2011), although one would anticipate that long stretches (especially as intact ORFs) would almost always be disrupted. Whether free DNA from the environment is assimilated by the host and made available to the virus during replication is an open question. Should intact non-host genes be present within infected *A. anophagefferens*, mechanisms like strand invasion might contribute to the assimilation of these foreign genes during replication of AaV inside the host (Filee et al., 2007).

Conclusions

We have presented the complete genome sequence of AaV, a virus that infects a marine pelagophyte that causes harmful brown tides. AaV is a large dsDNA virus, yet the smallest described member of an emerging clade, the Megaviridae, which harbors nucleocytoplasmic large DNA viruses with diverse hosts including both non-photosynthetic protists and photosynthetic algae. Despite having a smaller genome size, AaV shares a large number of core genes with other members of this growing group, which points to a common evolutionary history of AaV and some of the largest and most complex viruses. This observation suggests genome size is not a definitive criterion for the proposed 'Megaviridae' family. Analysis of the genome architecture suggests that the ancestral virus of AaV was probably much smaller in terms of

genome size, and likely followed an evolutionary pathway involving massive gene accumulation and gene duplication from the host as well as other organisms. The genome of AaV harbors putative functional and hypothetical genes from both *Phycodna-* and *Mimiviridae* clades and thereby enhances our understanding of the evolutionary history of these two diverged families which may have a common ancestor (Iyer et al., 2006). Moreover, AaV has several genes novel to the NCLDV group with possible roles in regulation of host cell processes. Several genes likely acquired from *A. anophagefferens* possibly allow AaV to facilitate the acquisition of resources by its host during viral infection.

Most of the Megaviridae family members isolated so far infect one single host, *Acanthamoeba polyphaga*, a non-photosynthetic phagotroph, although many researchers do not consider the amoeba to be the "native" host. AaV joins this group as one of two Megaviridae members having known photosynthetic hosts, and in the current case it brings a substantial ecological history (Gastrich et al., 2002, 2004; Gobler et al., 2007; Gobler and Sunda, 2012; Rowe et al., 2008). Algal viruses of the *Phycodnaviridae* family have been extensively studied regarding their seasonal dynamics (Martínez et al., 2007), diversity (reviewed in Short, 2012) and capability of modulating host cellular processes (Vardi et al., 2012) – however it is evident that algae-infecting NCLDVs exist across at least two distinct phylogenetic clades, and we anticipate new Megaviridae members with photosynthetic hosts will be described in near future. Because of the socioeconomic and environmental impact of brown tides, both *A. anophagefferens* and AaV have been studied extensively from physiological and ecological perspectives. Now, with the availability of genome sequences for both AaV and its host (Gobler et al., 2011), it is possible to develop a biomolecular experimental model system for teasing apart not only the dynamics of Megaviridae family, as well as to begin to experimentally gain insight into the genomic and phylogenetic evolution of this group.

Materials and methods

AaV production and purification

The original AaV virus was isolated in 2002 and has been maintained in culture since (see Rowe et al., 2008 for a description of the morphology and infectious potential of the particle). AaV was added at a ratio of 8–12 viruses/cell to a 650 mL culture of *A. anophagefferens* CCMP1984 at 18 °C, 14:10 light/dark cycle. The culture was monitored for virus and cell numbers. During lysis and at each step of purification, 40 μL of lysate was fixed with glutaraldehyde (0.5% final) for 15 min at 4 °C, then stored at −80 °C until staining and counting. Lysate was filtered through GF/F filters (Whatman) to remove debris and the majority of bacteria. Viral enrichment was performed using precipitation with PEG8000 (Lawrence and Steward, 2010). PEG8000 was added to filtered lysate (8 g per 100 mL of lysate) and completely dissolved by gentle mixing. The PEG/lysate solution was left overnight at 4 °C, followed by centrifugation for 35 min at 4 °C, 10,000*g*. Supernatants were carefully decanted, leaving 3–4 mL of supernatant in each bottle, residual liquid used to thoroughly rinse the bottles and the contents pooled. Approximately 10 mL of this concentrated virus solution was further concentrated to a volume of 1.5 mL using a 30 kDa cutoff Centricon filter (Millipore). Viruses were purified on an Optiprep™ (Iodixanol) step gradient. Four steps, 25%, 30%, 35% and 40%, were prepared by diluting the 60% Optiprep stock with MilliQ H_2O. 2.63 mL of each concentration was bottom loaded in a 12 mL ultracentrifuge tube, the lightest added first and the heaviest last (Lawrence and Steward, 2010). The 1.5 mL sample was then loaded on the top of the gradient. An identical gradient was prepared as a balance. The gradient was centrifuged in an SW41 rotor (Beckman Coulter) for 14 h and 45 min at 39,000 rpm. Starting from the top of the gradient and working down, 14 fractions of 0.6–1.0 mL were collected. The density of each fraction was determined. A 5 μl sample of each fraction was diluted into 995 μl 0.22 μm filtered media. 40 μl of each diluted sample was fixed with 0.8 μl glutaraldehyde for 15 min and viruses were enumerated by flow cytometry after staining wtith SYBR gold (Brussaard, 2004). Bacterial concentrations were determined simultaneously with virus counts. Fractions with the highest concentrations of virus and lowest concentrations of bacteria were pooled for extraction of the viral genomic DNA.

AaV DNA extraction

The extraction protocol closely followed that for the *E. huxleyi* virus (Schroeder et al., 2002). Briefly, the sample was treated with 5 mg/mL proteinase K in a lysis buffer consisting of 20 mM EDTA (pH 8.0) and 0.5% SDS at 65 °C, heated for approximately 1 h to break up the capsid. 10% of the sample volume of phenol was then added and the DNA was extracted with chloroform:isoamyl alcohol (CIA). Additional chloroform:isoamyl alcohol (24:1) and precipitation steps were inserted to reduce spectral interference from iodixanol, allowing DNA quantification by a NanoDrop™ 1000 (Thermo Scientific).

Genome assembly

The AaV genome was sequenced at an extremely high depth using Illumina™ technology and was further complemented by 454™ sequencing. 128,117,014 Illumina™ paired-end reads (256,234,029 total) of 100 bp length and 115,372 454™ single reads (avg. length of 272 bp) were used for assembling the genome of AaV. After removing the Illumina™ and 454™ specific adapters and trimming these reads based on quality scores (limit 0.04), a hybrid *de novo* assembly was performed on these reads in *CLC Genomics Workbench* 5.0 (www.clcbio.com) (paired distance range 120–400, K-mer size 64) resulting in 185,000 contigs with a 200 bp contig size cut-off. However, when a 2000 bp size cut-off was imposed, number of contigs was reduced to 136. The largest contig obtained was 121.7 kbp in length. Preliminary homology search of this 2000 bp subset against a local NCBI nr database (Benson et al., 2005) using BLASTX algorithm (Gish and States, 1993) revealed that several of these contigs originated from host genome, mitochondria and chloroplast sequences, indicating contamination from these sources during DNA extraction step and were not studied further. To identify the contigs of viral origin, we performed tBLASTx analysis (*E*-value < 1e-5) of these contigs against all the NCLDV virus genomes available as of November 21, 2012. Seven large contigs (A: 121,756 bp, B: 70,494 bp, C: 59,347 bp, D: 32,251 bp, E: 31,234 bp, F: 28,705 bp and G: 25,370 bp) were found to be putatively of viral origin based on the observation that they had highest sequence similarity to large dsDNA viruses with best hits to different sequenced NCLDVs. Sequence statistics analysis revealed that six of these seven putative contigs had very similar GC content ranging from 27.7% (contig 'D') to 29.4% (contig 'F'). Among all the 136 contigs analyzed, these were the only ones having GC contents below 30% whereas contig 'G' had a slightly higher GC content of 31.4%.

Assuming that gaps between these contigs of viral origin resulted from repetitive sequences, we performed an all-vs-all nucleotide BLAST of these seven contigs in search of shared sequence similarities at the ends. This analysis revealed that contig pair (G,F); (F,A); (A,C); (C,D); (D,E); (E,B); (E,C) and (B,F) indeed shared sequence similarity at their ends. We hypothesized that these contig pairs represent contiguous sequences, and designed

forward and reverse primers for the corresponding pairs. The PCR products spanning the gaps were purified and cloned using TOPO® TA cloning® kit in One Shot® TOPO10 chemically competent *Escherichia coli* cells (Invitrogen™, CA) and sequenced using Sanger method. When the gaps were large, a second round of primers were designed from the ends of the sequences obtained from first round of reactions. The generated sequences were used to close the gaps between these contigs to obtain a continuous sequence of 370,920 bp, which represents the genome of AaV.

Contig 'G' (25,370 bp) represents the 5′ terminal region of the genome, which consists of 20 DUF285 domain-containing CDSs in tandem (interspersed by three other CDSs). This contig also has a slightly higher GC% (31.4%) compared to the other viral contigs. As discussed in the main text, this region is also highly repetitious. To verify that this conting was not an artifact of misassembled reads, we designed seven different sets of primers spanning seven non-overlapping regions of Contig 'G'. Upon PCR amplification, each of these primer pairs generated amplicons of the expected size, which confirmed that Contig 'G' is not a product of misassembled reads due to its highly repetitious nature, but an authentic part of the AaV genome. As part of our ongoing genome analysis, several other genes (including the DNA polymerase, major capsid, putative ion channel and mechanosensitive channel) have already been verified by cloning and Sanger sequencing. The genome sequence data of AaV has been deposited in the GenBank database (accession no. KJ645900).

Genome annotation

Putative coding sequences (CDS) were predicted using Prodigal web server (Hyatt et al., 2010). We defined a CDS having a minimum length of 50 consecutive codons bordered by a start and a stop codon. Homologous genes for the determined CDSs were detected by carrying out BLASTp analysis (Altschul et al., 1997) of the CDSs against the NCBI non-redundant protein database (nr) (Benson et al., 2005) with an *E*-value cutoff of 1e-05 to avoid false positive matches. Protein domains were detected using NCBI Conserved Domain Database (CDD) (Marchler-Bauer et al., 2013), pfam (Punta et al., 2012) and Interpro (Quevillon et al., 2005) servers. Functional annotation resulted from integrating BLASTp results with the results obtained from these databases. tRNAs were predicted in the tRNAscanSE (Schattner et al., 2005) server using the general tRNA model.

Phylogenetic analysis

Putative homologs of the query proteins were identified by separate BLASTp (Altschul et al., 1997) searches against the viruses, eukaryotes, bacteria and other taxonomic subgroups defined in the GenBank nr database (Benson et al., 2005). For each of the query proteins, a representative set of homologs were selected. When available, homologs from the host algae were included in the phylogenetic reconstructions. Multiple sequence alignments were performed in MEGA 5.0 software (Tamura et al., 2011) using MUSCLE algorithm (Edgar, 2004) followed by manual refinement. Evolutionary models having the highest likelihood for each set of alignments were determined using Prottest 3.2.1 (Darriba et al., 2011). Maximum Likelihood phylogenetic reconstruction was performed for each set of alignments in TREEFINDER (Jobb et al., 2004). The Expected-Likelihood Weights (ELW) of 1000 local rearrangements were used as confidence values for the nodes.

A number of CDSs had no homologs outside a particular domain of life or outside NCLDVs (NCLDV specific hypothetical CDSs). Phylogenetic origins of these CDSs were assigned directly to the respective domains of life they had best match to (*E*-value < 1e-05).

Other analyses

BLASTn comparison of the whole genomes was performed and illustrated using BLAST Ring Image Generator (BRIG) (Alikhan et al., 2011) with an *E*-value cutoff of 1e-05. For NCVOG analysis, a BLASTp (Altschul et al., 1997) search of AaV CDSs was carried out against a database containing all NCLDV proteins belonging to different NCVOG categories as constructed by Yutin et al. (2009) (downloaded from ftp://ftp.ncbi.nih.gov/pub/wolf/COGs/NCVOG). Hits with E-values lower than 1e-05 were assigned to their representative NCVOGs. Conserved motifs upstream of the PARCELs were analyzed using the MEME suite (Bailey and Elkan, 1994). Whole genome dot plot was constructed in YASS (Yet Another Similarity Searcher) server (Noé and Kucherov, 2005).

Acknowledgments

The authors thank a decade of students and collaborators for their continued support of work with the *Aureococcus anophagefferens* system. Isolation and purification efforts of AaVs and genomic DNA was supported by the National Science Foundation (OCE-1061883 to KDB) and an Institute of Marine and Coastal Sciences postdoctoral fellowship to CMB. Sequencing, assembly and annotation were also supported by the National Science Foundation (EF-0949120 and OCE-1061352 to SWW, EF-0949162 to WHW).

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.virol.2014.06.031.

References

Akiyama, M., Maki, H., Sekiguchi, M., Horiuchi, T., 1989. A specific role of MutT protein: to prevent dG.dA mispairing in DNA replication. Proc. Natl. Acad. Sci. USA 86, 3949–3952.

Alikhan, N.-F., Petty, N., Ben Zakour, N., Beatson, S., 2011. BLAST Ring Image Generator (BRIG): simple prokaryote genome comparisons. BMC Genomics 12, 402.

Allen, M.J., Schroeder, D.C., Wilson, W.H., 2006. Preliminary characterisation of repeat families in the genome of EhV-86, a giant algal virus that infects the marine microalga *Emiliania huxleyi*. Arch. Virol. 151, 525–535.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.

Bailey, T.L., Elkan, C., 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. Int. Conf. Intell. Syst. Mol. Biol. 2, 28–36.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Wheeler, D.L., 2005. GenBank. Nucleic Acids Res. 33, D34–D38.

Bidle, K.D., Haramaty, L., Barcelos e Ramos, J., Falkowski, P., 2007. Viral activation and recruitment of metacaspases in the unicellular coccolithophore, *Emiliania huxleyi*. Proc. Natl. Acad. Sci. USA 104, 6049–6054.

Bidle, K.D., Vardi, A., 2011. A chemical arms race at sea mediates algal host–virus interactions. Curr. Opin. Microbiol. 14, 449–457.

Boyer, M., Madoui, M.-A., Gimenez, G., La Scola, B., Raoult, D., 2010. Phylogenetic and phyletic studies of informational genes in genomes highlight existence of a 4th domain of life including giant viruses. PLoS ONE 5, e15530.

Boyer, M., Yutin, N., Pagnier, I., Barrassi, L., Fournous, G., Espinosa, L., Robert, C., Azza, S., Sun, S., Rossmann, M.G., Suzan-Monti, M., La Scola, B., Koonin, E.V., Raoult, D., 2009. Giant Marseillevirus highlights the role of amoebae as a melting pot in emergence of chimeric microorganisms. Proc. Natl. Acad. Sci. USA 107 (45), 19508–19513.

Brussaard, C.P., 2004. Optimization of procedures for counting viruses by flow cytometry. Appl. Environ. Microbiol. 70, 1506–1513.

Darriba, D., Taboada, G.L., Doallo, R., Posada, D., 2011. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics 27 (8), 1164–1165.

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797.

Filee, J., Pouget, N., Chandler, M., 2008. Phylogenetic evidence for extensive lateral acquisition of cellular genes by Nucleocytoplasmic large DNA viruses. BMC Evol. Biol. 8, 320.

Filee, J., Siguier, P., Chandler, M., 2007. I am what I eat and I eat what I am: acquisition of bacterial genes by giant viruses. Trends Genet. 23, 10–15.

Fischer, M.G., Allen, M.J., Wilson, W.H., Suttle, C.A., 2010. Giant virus with a remarkable complement of genes infects marine zooplankton. Proc. Natl. Acad. Sci. USA 107 (45), 19508–19513.
Fitzgerald, L.A., Graves, M.V., Li, X., Hartigan, J., Pfitzner, A.J., Hoffart, E., Van Etten, J. L., 2007. Sequence and annotation of the 288-kb ATCV-1 virus that infects an endosymbiotic Chlorella strain of the heliozoon *Acanthocystis turfacea*. Virology 362, 350–361.
Garry, R.T., Hearing, P., Cosper, E.M., 1998. Characterization of a lytic virus infectious to the bloom-forming microalga *Aureococcus anophagefferens* (Pelagophyceae). J. Phycol. 34, 616–621.
Gastrich, M., Anderson, O.R., Cosper, E., 2002. Viral-like particles (VLPS) in the alga, *Aureococcus anophagefferens* (Pelagophyceae), during 1999–2000 Brown tide blooms in Little Egg Harbor, New Jersey. Estuaries 25, 938–943.
Gastrich, M., Leigh-Bell, J., Gobler, C., Roger Anderson, O., Wilhelm, S., Bryan, M., 2004. Viruses as potential regulators of regional brown tide blooms caused by the alga, *Aureococcus anophagefferens*. Estuaries 27, 112–119.
Gastrich, M.D., Anderson, O.R., Benmayor, S.S., Cosper, E.M., 1998. Ultrastructural analysis of viral infection in the brown-tide alga, *Aureococcus anophagefferens* (Pelagophyceae). Phycologia 37, 300–306.
Gish, W., States, D.J., 1993. Identification of protein coding regions by database similarity search. Nat. Genet. 3, 266–272.
Gobler, C.J., Anderson, O.R., Gastrich, M.D., Wilhelm, S.W., 2007. Ecological aspects of viral infection and lysis in the harmful brown tide alga *Aureococcus anophagefferens*. Aquat. Microb. Ecol. 47, 25–36.
Gobler, C.J., Berry, D.L., Dyhrman, S.T., Wilhelm, S.W., Salamov, A., Lobanov, A.V., Zhang, Y., Collier, J.L., Wurch, L.L., Kustka, A.B., Dill, B.D., Shah, M., VerBerkmoes, N.C., Kuo, A., Terry, A., Pangilinan, J., Lindquist, E.A., Lucas, S., Paulsen, I.T., Hattenrath-Lehmann, T.K., Talmage, S.C., Walker, E.A., Koch, F., Burson, A.M., Marcoval, M.A., Tang, Y.-Z., LeCleir, G.R., Coyne, K.J., Berg, G.M., Bertrand, E.M., Saito, M.A., Gladyshev, V.N., Grigoriev, I.V., 2011. Niche of harmful alga *Aureococcus anophagefferens* revealed through ecogenomics. Proc. Natl. Acad. Sci. USA 08 (11), 4352–4357.
Gobler, C.J., Sunda, W.G., 2012. Ecosystem disruptive algal blooms of the brown tide species, *Aureococcus anophagefferens* and *Aureoumbra lagunensis*. Harmful Algae 14, 36–45.
Greiner, T., Frohns, F., Kang, M., Van Etten, J.L., Kasmann, A., Moroni, A., Hertel, B., Thiel, G., 2009. Chlorella viruses prevent multiple infections by depolarizing the host membrane. J. Gen. Virol. 90, 2033–2039.
Hess, D.C., Myers, C.L., Huttenhower, C., Hibbs, M.A., Hayes, A.P., Paw, J., Clore, J.J., Mendoza, R.M., Luis, B.S., Nislow, C., Giaever, G., Costanzo, M., Troyanskaya, O. G., Caudy, A.A., 2009. Computationally driven, quantitative experiments discover genes required for mitochondrial biogenesis. PLoS Genet. 5, e1000407.
Hill, E., 2006. The cyanophage molecular mixing bowl of photosynthesis genes. PLoS Biol. 4, e264.
Hyatt, D., Chen, G.-L., LoCascio, P., Land, M., Larimer, F., Hauser, L., 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinform. 11, 119.
Iyer, L.M., Balaji, S., Koonin, E.V., Aravind, L., 2006. Evolutionary genomics of nucleocytoplasmic large DNA viruses. Virus Res. 117, 156–184.
Jobb, G., von Haeseler, A., Strimmer, K., 2004. TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. BMC Evol. Biol. 4, 18.
Kislyuk, A., Haegeman, B., Bergman, N., Weitz, J., 2011. Genomic fluidity: an integrative view of gene diversity within microbial populations. BMC Genomics 12, 32.
Koonin, E.V., Yutin, N., 2010. Origin and evolution of eukaryotic large nucleocytoplasmic DNA viruses. Intervirology 53, 284–292.
Lawrence, J.E., Steward, G.F., 2010. Purification of viruses by centrifugation. In: Wilhelm, S.W., Weinbauer, M.G., Suttle, C.A. (Eds.), Manual of Aquatic Viral Ecology (MAVE). Association for the Sciences of Limnology and Oceanography, Texas, pp. 166–181.
Legendre, M., Bartoli, J., Shmakova, L., Jeudy, S., Labadie, K., Adrait, A., Lescot, M., Poirot, O., Bertaux, L., Bruley, C., Couté, Y., Rivkina, E., Abergel, C., Claverie, J.-M., 2014. Thirty-thousand-year-old distant relative of giant icosahedral DNA viruses with a Pandoravirus morphology. Proc. Natl. Acad. Sci. USA, http://dx.doi.org/10.1073/pnas.1320670111 (Epub ahead of print).
Liu, S.-L., Baute, G.J., Adams, K.L., 2011. Organ and cell type-specific complementary expression patterns and regulatory neofunctionalization between duplicated genes in *Arabidopsis thaliana*. Genome Biol. Evol. 3, 1419–1436.
Marchler-Bauer, A., Zheng, C., Chitsaz, F., Derbyshire, M.K., Geer, L.Y., Geer, R.C., Gonzales, N.R., Gwadz, M., Hurwitz, D.I., Lanczycki, C.J., Lu, F., Lu, S., Marchler, G. H., Song, J.S., Thanki, N., Yamashita, R.A., Zhang, D., Bryant, S.H., 2013. CDD: conserved domains and protein three-dimensional structure. Nucleic Acids Res. 41, D348–D352.
Marmorstein, R., Roth, S.Y., 2001. Histone acetyltransferases: function, structure, and catalysis. Curr. Opin. Genet. Dev. 11, 155–161.
Martínez, J.M., Schroeder, D.C., Larsen, A., Bratbak, G., Wilson, W.H., 2007. Molecular dynamics of *Emiliania huxleyi* and cooccurring viruses during two separate mesocosm studies. Appl. Environ. Microbiol. 73, 554–562.
Milligan, K.L., Cosper, E.M., 1994. Isolation of virus capable of lysing the brown tide microalga, *Aureococcus anophagefferens*. Science 266, 805–807.
Monier, A., Pagarete, A., de Vargas, C., Allen, M.J., Read, B., Claverie, J.M., Ogata, H., 2009. Horizontal gene transfer of an entire metabolic pathway between a eukaryotic alga and its DNA virus. Genome Res. 19, 1441–1449.
Moreau, H., Piganeau, G., Desdevises, Y., Cooke, R., Derelle, E., Grimsley, N., 2010. Marine prasinovirus genomes show low evolutionary divergence and acquisition of protein metabolism genes by horizontal gene transfer. J. Virol. 84, 12555–12563.
Nemova, N.N., Lysenko, L.A., Kantserova, N.P., 2010. Proteases of the calpain family: structure and functions. Russ. J. Dev. Biol. 41, 318–325.
Neupartl, M., et al., 2008. Chlorella viruses evoke a rapid release of K+ from host cells during the early phase of infection. Virology 372 (2), 340–348.
Noé, L., Kucherov, G., 2005. YASS: enhancing the sensitivity of DNA similarity search. Nucleic Acids Res. 33, W540–W543.
Ogata, H., Ray, J., Toyoda, K., Sandaa, R.A., Nagasaki, K., Bratbak, G., Claverie, J.M., 2011. Two new subfamilies of DNA mismatch repair proteins (MutS) specifically abundant in the marine environment. ISME J. 5, 1143–1151.
Ohi, M.D., Vander Kooi, C.W., Rosenberg, J.A., Chazin, W.J., Gould, K.L., 2003. Structural insights into the U-box, a domain associated with multi-ubiquitination. Nat. Struct. Mol. Biol. 10, 250–255.
Pruzinska, A., Tanner, G., Anders, I., Roca, M., Hortensteiner, S., 2003. Chlorophyll breakdown: pheophorbide a oxygenase is a Rieske-type iron-sulfur protein, encoded by the accelerated cell death 1 gene. Proc. Natl. Acad. Sci. USA 100, 15259–15264.
Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E.L.L., Eddy, S.R., Bateman, A., Finn, R.D., 2012. The Pfam protein families database. Nucleic Acids Res. 40, D290–D301.
Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., Lopez, R., 2005. InterProScan: protein domains identifier. Nucleic Acids Res. 33, W116–W120.
Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., La Scola, B., Suzan, M., Claverie, J.-M., 2004. The 1.2-megabase genome sequence of Mimivirus. Science 306, 1344–1350.
Rose, S.L., Fulton, J.M., Brown, C.M., Natale, F., Van Mooy, B.A.S., Bidle, K.D., 2014. Isolation and characterization of lipid rafts in *Emiliania huxleyi*: a role for membrane microdomains in host–virus interactions. Environ. Microbiol. 16 (4), 1150–1166.
Roske, K., Foecking, M., Yooseph, S., Glass, J., Calcutt, M., Wise, K., 2010. A versatile palindromic amphipathic repeat coding sequence horizontally distributed among diverse bacterial and eukaryotic microbes. BMC Genomics 11, 430 (s).
Rowe, J.M., Dunlap, J.R., Gobler, C.J., Anderson, O.R., Gastrich, M.D., Wilhelm, S.W., 2008. Isolation of a non-phage-like lytic virus infecting *Aureococcus anophagefferens*. J. Phycol. 44, 71–76.
Santini, S., Jeudy, S., Bartoli, J., Poirot, O., Lescot, M., Abergel, C., Barbe, V., Wommack, K.E., Noordeloos, A.A.M., Brussaard, C.P.D., Claverie, J.-M., 2013. Genome of *Phaeocystis globosa* virus PgV-16T highlights the common ancestry of the largest known DNA viruses infecting eukaryotes. Proc. Natl. Acad. Sci. USA 110 (26), 10800.
Schattner, P., Brooks, A.N., Lowe, T.M., 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. 33, W686–W689.
Schroeder, D.C., Oke, J., Malin, G., Wilson, W.H., 2002. Coccolithovirus (Phycodnaviridae): characterisation of a new large dsDNA algal virus that infects *Emiliana huxleyi*. Arch. Virol. 147, 1685–1698.
Short, S.M., 2012. The ecology of viruses that infect eukaryotic algae. Environ. Microbiol. 14, 2253–2271.
Sieburth, J.M., Johnson, P.W., Hargraves, P.E., 1988. Ultrastructure and ecology of *Aureococcus anophageferens* gen. et sp. nov. (Chrysophyceae): the dominant picoplankter during a bloom in Narragansett Bay, Rhode Island, summer 19851. J. Phycol. 24, 416–425.
Simonin, Y., Disson, O., Lerat, H., Antoine, E., Biname, F., Rosenberg, A.R., Desagher, S., Lassus, P., Bioulac-Sage, P., Hibner, U., 2009. Calpain activation by Hepatitis C virus proteins inhibits the extrinsic apoptotic signaling pathway. Hepatology 50, 1370–1379.
Srinivasan, V., Schnitzlein, W.M., Tripathy, D.N., 2001. Fowlpox virus encodes a novel DNA repair enzyme, CPD-photolyase, that restores infectivity of UV light-damaged virus. J. Virol. 75, 1681–1688.
Suhre, K., 2005. Gene and genome duplication in *Acanthamoeba polyphaga* Mimivirus. J. Virol. 79, 14095–14101.
Szkopinska, A., 2000. Ubiquinone. Biosynthesis of quinone ring and its isoprenoid side chain. Intracellular localization. Acta Biochim. Pol. 47, 469–480.
Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 28, 2731–2739.
Thiel, G., Baumeister, D., Schroeder, I., Kast, S.M., Van Etten, J.L., Moroni, A., 2011. Minimal art: or why small viral K+ channels are good tools for understanding basic structure and function relations. Biochim. Biophys. Acta 1808, 580–588.
Vardi, A., Haramaty, L., Van Mooy, B.A., Fredricks, H.F., Kimmance, S.A., Larsen, A., Bidle, K.D., 2012. Host–virus dynamics and subcellular controls of cell fate in a natural coccolithophore population. Proc. Natl. Acad. Sci. USA 109, 19327–19332.
Vardi, A., Van Mooy, B.A.S., Fredricks, H.F., Popendorf, K.J., Ossolinski, J.E., Haramaty, L., Bidle, K.D., 2009. Viral glycosphingolipids induce lytic infection and cell death in marine phytoplankton. Science 326, 861–865.
Weinbauer, M.G., Wilhelm, S.W., Suttle, C.A., Garza, D.R., 1997. Photoreactivation compensates for UV damage and restores infectivity to natural marine virus communities. Appl. Environ. Microbiol. 63, 2200–2205.
Williams, T.A., Embley, T.M., Heinz, E., 2011. Informational gene phylogenies do not support a fourth domain of life for Nucleocytoplasmic Large DNA viruses. PLoS ONE 6, e21080.
Wilson, M.E., Maksaev, G., Haswell, E.S., 2013. MSCS-like mechanosensitive channels in plants and microbes. Biochemistry 52, 5708–5722.

Wilson, W.H., Schroeder, D.C., Allen, M.J., Holden, M.T.G., Parkhill, J., Barrell, B.G., Churcher, C., Hamlin, N., Mungall, K., Norbertczak, H., Quail, M.A., Price, C., Rabbinowitsch, E., Walker, D., Craigon, M., Roy, D., Ghazal, P., 2005. Complete genome sequence and lytic phase transcription profile of a coccolithovirus. Science 309, 1090–1092.

Yau, S., Lauro, F.M., DeMaere, M.Z., Brown, M.V., Thomas, T., Raftery, M.J., Andrews-Pfannkoch, C., Lewis, M., Hoffman, J.M., Gibson, J.A., Cavicchioli, R., 2011. Virophage control of Antarctic algal host–virus dynamics. Proc. Natl. Acad. Sci. USA 108, 6163–6168.

Yutin, N., Colson, P., Raoult, D., Koonin, E.V., 2013. Mimiviridae: clusters of orthologous genes, reconstruction of gene repertoire evolution and proposed expansion of the giant virus family. Virol. J. 10, 106.

Yutin, N., Wolf, Y., Raoult, D., Koonin, E., 2009. Eukaryotic large nucleo-cytoplasmic DNA viruses: clusters of orthologous genes and reconstruction of viral genome evolution. Virol. J. 6, 223.

Zhang, Q.-C., Qiu, L.-M., Yu, R.-C., Kong, F.-Z., Wang, Y.-F., Yan, T., Gobler, C.J., Zhou, M.-J., 2012. Emergence of brown tides caused by *Aureococcus anophagefferens* Hargraves et Sieburth in China. Harmful Algae 19, 117–124.

Zhu, W., Li, J., Liang, G., 2011. How does cellular heparan sulfate function in viral pathogenicity? Biomed. Environ. Sci. 24, 81–87.